Versión 1.0: 22/08/2014
Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Guía de catalogación de datos abiertos

Hoy en día los datos constituyen recursos esenciales con un enorme potencial para la generación de beneficios tanto sociales como económicos. Muchos de esos valiosos **conjuntos de datos** son gestionados por la Administración, y el objetivo que se persigue con los **datos abiertos** (Open Data) es la publicación de esos datos de forma que puedan ser reutilizados para que todo el mundo se pueda beneficiar de ellos.

Sin embargo, la cantidad y variedad de datos que se producen dentro de cualquier Administración hace que esta no sea una tarea trivial y que requiera de la colaboración de todas las áreas y organismos que la componen.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Antecedentes

El Gobierno de Aragón, mediante acuerdo de Gobierno de 17 de Julio de 2012 se compromete a la efectiva apertura de los datos públicos que obran en su poder siguiendo las pautas de la Directiva 2003/98/CE sobre Reutilización de la Información del Sector Público y su trasposición a nivel nacional en España mediante la Ley 37/2007. Para ello se pone en marcha el proyecto Aragón Open Data, cuya finalidad será proporcionar la plataforma tecnológica básica que de soporte a la publicación de los datos en abierto.

Por otra parte a fecha 4 de marzo de 2013 se publica en el BOE la norma técnica de interoperabilidad de reutilización de recursos de la información en España, cuyo objetivo es facilitar y garantizar el proceso de reutilización de la información de las Administraciones públicas en todos los niveles, asegurando la persistencia de la información, el uso de los formatos adecuados, así como los términos y condiciones de uso.

La publicación de esta norma supone un hito importante, ya que en ella se describen detalladamente los metadatos a utilizar para garantizar el descubrimiento y la interoperabilidad de los recursos de información elaborados o custodiados por el sector público. Sin embargo, para que esto sea posible es necesario ser capaz de aplicar adecuada y **homogéneamente** esos metadatos de forma que se eviten ambigüedades en su interpretación. Este hecho es fundamental para que el tratamiento automatizado de la información sea no sólo técnicamente posible, sino también semánticamente fiable.

Este documento es una guía destinada a los responsables de los datos dentro de la Administración para servir como ayuda y orientación en el proceso de catalogación de los datos que gestionan, así como para la correcta interpretación y aplicación de los metadatos a través de indicaciones claras y ejemplos de su uso.

El objetivo de esta guía es ser distribuida entre los distintos organismos de la administración de forma que les proporcione las indicaciones necesarias para elaborar, gestionar y mantener toda la información relativa a los metadatos de los conjuntos de datos de forma coherente y conforme a la norma, pudiendo de este modo auto-gestionarse utilizando las herramientas que el proyecto habrá puesto previamente a su disposición a través de la plataforma Aragón Open Data.

Los principios de los datos abiertos

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Siempre que publicamos datos en abierto debemos seguir los principios que rigen el *Open Goverment Data*[1] para garantizar que se haga de forma efectiva:

- **Completos**: Deben proporcionarse todos los datos con la única excepción de aquellos que cuenten con especial protección por cuestiones de privacidad, seguridad o similares.
- **Primarios**: Los datos deben compartirse tal y como se originan y con el mismo nivel de detalle, sin modificaciones, agregados u otras operaciones intermedias adicionales.
- **En tiempo**: Los datos se proporcionan en el momento adecuado para que no pierdan su valor por quedar obsoletos.
- **Accesibles**: Los datos están disponibles de forma igualitaria y equitativa para todos los usuarios.
- **Legibles por máquinas**: Los datos estarán estructurados de forma que permitan su tratamiento automatizado.
- **No discriminatorios**: El acceso a los datos será abierto e inmediato bajo demanda sin requisitos adicionales de registro o identificación.
- **Formatos libres**: Los datos se publican en formatos no propietarios para maximizar su potencial de reutilización.
- **Licencias abiertas**: Los datos se comparten libres de derechos, de forma que cualquiera pueda usarlos, reutilizarlos y redistribuirlos sin restricciones.

A lo largo de esta guía se explicarán una serie de pautas que son de utilidad a la hora de conseguir cumplir con estos principios.

Definiciones

A continuación se incluyen las definiciones de algunos conceptos clave relacionados con el mundo de los datos abiertos que serán utilizados a lo largo del documento.

Catálogo de datos: repositorio electrónico donde se almacenan y administran datos y documentos electrónicos, junto a sus metadatos.

Conjunto de datos: serie de datos asociados en torno a una temática o materia que los caracteriza y que puede estar compuesto por una o más *distribuciones* normalmente en diferentes formatos.

[1] Open Government Data Principles: https://public.resource.org/8_principles.html

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Dato: representación mínima de unos hechos, conceptos o instrucciones de un modo formalizado, y adecuado para su comunicación, interpretación o procesamiento por medios automáticos o humanos.

Datos abiertos (Open Data): iniciativa de apertura y compartición de los datos de forma apta para su reutilización por parte de terceros.

Datos personales: toda información de cualquier tipo que permita la identificación de una persona física, ya sea de forma directa o indirecta.

Distribución: datos disponibles en un formato concreto de presentación, ya sea un tipo de fichero, un servicio web u otro. Un mismo conjunto de datos puede disponer de múltiples distribuciones en distintos formatos para facilitar su reutilización con distintos programas y herramientas, dependiendo del perfil del destinatario o *reutilizador* final.

Infomediario: agente que recolecta, organiza y agrega valor a grandes cantidades de datos provenientes de diversas fuentes, actuando así como intermediario entre los proveedores y los consumidores finales.

Metadatos: datos que definen y describen a otros datos. Existen diferentes tipos de metadatos según su aplicación.

Reutilizador: persona, física o jurídica que reutilice información del sector público, ya sea para fines comerciales o no comerciales, siempre que dicho uso no constituya una actividad administrativa pública.

Se pueden consultar otras definiciones y acrónimos relacionados con la Reutilización de la Información en el Sector Público en la guía publicada al respecto dentro del Esquema Nacional de Interoperabilidad[2].

El tratamiento de los datos

Para una gestión óptima de los datos de cara a su reutilización es necesario establecer una recopilación inicial de los mismos en un único punto de acceso común, el **catálogo de datos**, lo que permitirá un acceso más fácil por parte de los potenciales *reutilizadores*.

El catálogo de los conjuntos de datos se generará en base a la información y los *metadatos* proporcionados por los distintos responsables, que finalmente serán publicados a través de la plataforma online preparada por el Gobierno de Aragón

[2] Reutilización de Información del Sector Público - Definiciones y Acrónimos: http://datos.gob.es/sites/default/files/definiciones-acronimos_risp.docx

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

http://opendata.aragon.es/, que permitirá recopilar el inventario de todos los conjuntos de datos públicos reutilizables en la región.

El objetivo final es conseguir **catalogar** la información sobre los datos disponibles en la actualidad en todas las áreas y organismos del Gobierno de Aragón y otras entidades relacionadas, para lo que será necesario seguir una serie de pasos que se explican en las siguientes secciones:

1 – LOCALIZACIÓN DE LOS DATOS

Como punto de partida inicial debería llevarse a cabo un proceso de inventariado de datos para conocer cuál es el alcance y extensión de los registros de datos existentes. Ese inventario inicial debe considerarse tan solo como el punto de partida inicial de un proceso que deberá **expandirse progresivamente hasta incluir todas las áreas y organismos**. Además, dado que la producción y actualización de datos es continua, este proceso debe verse también como algo **cíclico y continuo** dentro de cada área y organismo donde se deberá repetir y hacer perdurable a lo largo del tiempo.

En general, los gestores de los datos deben asumir y asimilar como responsabilidad propia tanto la identificación inicial de los conjuntos de datos existentes como la posterior detección de nuevos conjuntos de datos, así como de los cambios y actualizaciones que se vayan produciendo en los mismos.

Algunos momentos durante el transcurso de la actividad diaria de cualquier área u organismo en los que será más fácil que se produzcan nuevos conjuntos de datos son:

- Cuando se inicia una **nueva actividad o proyecto**;
- Cuando se introducen cambios significativos que afecten a las tecnologías y/o los **sistemas de información** utilizados;
- Cuando sea necesaria información adicional para fundamentar una **toma de decisiones**, ya sea ante cambios previamente planificados o por nuevas situaciones no previstas;
- Cuando aparecen **nuevas necesidades de información** demandadas desde otras áreas u organismos;
- Cuando se producen **cambios legislativos** que afecten al ámbito y objeto de actuación;
- Cuando se reciben nuevas **peticiones de datos** e información por parte de la ciudadanía u otros agentes *reutilizadores*;

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Aparte de todo lo anterior, existen también otros factores que pueden dar lugar a la generación o actualización masiva de datos y que dependerán de los procedimientos específicos establecidos en cada área u organismo.

Por otro lado, cuando los datos están disponibles a través de medios no electrónicos – como por ejemplo publicaciones impresas – o mediante soportes físicos – como CDs; DVDs; u otros – será necesario localizar la fuente original de los datos en las bases de datos y sistemas internos, con el fin de poder extraer el correspondiente conjunto de datos en formato electrónico de cara a su publicación.

A medio y largo plazo debe plantearse también la digitalización de cualquier contenido que todavía no esté disponible en un formato electrónico apropiado de cara a su posible reutilización.

2 – SELECCIÓN DE LOS DATOS PARA LA CATALOGACIÓN

El objetivo a largo plazo es conseguir **catalogar todos los datos disponibles**. Sin embargo, dado que el tiempo y los recursos disponibles siempre serán limitados, cada área u organismo, conocedor del tipo de información que posee y de su potencial audiencia, deberá determinar qué información es más valiosa desde el punto de vista de la *reutilización* para poder *priorizar* su catalogación.

Debido al carácter diverso de los organismos y de la información que gestionan la priorización se deberá adecuar a las particularidades de cada caso. Sin embargo, existen también algunos criterios generales que se deben valorar:

- **Normativa y legislación**, dado que cada vez es más frecuente la existencia de normativa y recomendaciones a nivel europeo, nacional o regional que aconseja, o incluso establece la obligación, publicar cierto tipo de información, ya sea relativa a transparencia u otras materias.
- **Relevancia de los datos para la ciudadanía**, valorando el impacto y la posible utilidad de los datos para la ciudadanía, especialmente en aquellos casos en los que ya se hayan recibido peticiones respecto a datos concretos.
- **Relevancia de los datos para las empresas**, valorando el impacto y la posible utilidad de los datos para las empresas que harán uso de ellos, especialmente en aquellos casos donde existan ya líneas de trabajo comunes en el ámbito público-privado.
- **Relevancia de los datos para la propia Administración y organismos relacionados**, valorando el valor estratégico y la posible utilidad de los datos para uso interno del propio Gobierno de Aragón o cualquier otro organismo público.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

- **Frecuencia de actualización de los datos** ya que, aunque no es una norma global y finalmente dependerá también de la naturaleza de la información, desde el punto de vista de la reutilización los datos más dinámicos y que cambian con mayor frecuencia ofrecen generalmente mayor valor y posibilidades de reutilización que los estáticos.
- **Volumen de datos contenidos en los conjuntos**, teniendo en cuenta tanto el volumen actual como las previsiones de crecimiento futuras ya que, como norma general, a mayor cantidad de datos más potencial de reutilización.

En cualquier caso, cada área u organismo podría también adaptar y completar su propia lista de criterios de forma acorde tanto a su ámbito de actividad y servicio como a la normativa particular vigente.

Restricciones en la publicación de datos

Aunque la norma general consiste simplemente en **publicar todo**, existen también ciertas limitaciones por causas de fuerza mayor a la hora de publicar datos que se deben respetar, como podría ser en los siguientes casos:

- Que por su **carácter personal** estén restringidos por la ley protección de datos de carácter personal;
- Que estén sometidos al **secreto estadístico** o a la **confidencialidad comercial**;
- Que afecten a la **defensa** o a la protección de la **seguridad pública**;
- Sobre los que exista **derecho de propiedad intelectual o industrial** por parte de terceros;
- Sobre los que exista **prohibición o limitación en el derecho de acceso**, o para su acceso se requiera ser titular de un derecho o interés legítimo;

En cualquiera de los casos anteriores la información sensible debe ser filtrada y extraída de los conjuntos de datos **previamente a su publicación**.

3 – PREPARACIÓN DE LOS DATOS

Uno de los principales objetivos a la hora de exponer los datos de carácter público es permitir y facilitar el **procesamiento automático de los datos**, ya que esta es la única manera de tratar grandes cantidades de datos de forma eficiente. Para que dicho procesamiento a través de programas informáticos sea posible es necesario que la información esté organizada en los **formatos adecuados** para que los datos puedan ser accedidos y tratada de forma automatizada.

Formatos

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Algunos criterios a seguir a la hora de seleccionar los formatos adecuados son:

1. Formatos **estructurados** que sean directamente **legibles por las máquinas**.
2. Los formatos que sean más demandados o populares entre los potenciales agentes *reutilizadores* o, en su defecto, aquellos **formatos de uso generalizado en los principales sectores de consumo de la información** y que por tanto supongan un estándar, ya sea oficial o *de facto*.
3. Formatos **abiertos** que permitan su uso **sin restricciones** de ningún tipo, ya sea en cuanto a derechos de autor o patentes, con lo que se favorecerá la igualdad de oportunidades e incluso un posible ahorro de costes para todas las partes implicadas. De lo contrario, aunque los datos fuesen publicados en abierto estaríamos obligados a emplear alguna tecnología específica ligada a un determinado proveedor para el acceso y/o posterior tratamiento de los mismos.
4. Formatos compatibles con lo establecido en la **Norma Técnica de Interoperabilidad del Catálogo de estándares**[3] siempre que sea posible, en particular en lo relativo a formatos de ficheros, gestión documental y archivística y semántica.
5. Formatos con los que se pueda ofrecer también *metainformación* sobre los esquemas o vocabularios utilizados para representar la información, como por ejemplo XML ó RDF[4].
6. Ofrecer una variedad de **formatos distintos y complementarios de forma simultánea**, siempre que sean adecuados para los datos en cuestión. De este modo será más fácil cubrir las necesidades de un mayor número de perfiles de *reutilizadores* con distintos nivel de conocimientos técnicos y preferencias (empresarios, emprendedores, investigadores, periodistas, etc.)

El objetivo ideal en cuanto a formatos sería conseguir que la publicación de los datos se realizase siempre de forma que **al menos una de las *distribuciones*** utilizase siempre un formato **abierto**, **estándar**, **estructurado** y **legible por las máquinas**. Ejemplos frecuentes de este tipo de formatos son: CSV (o TSV) para datos tabulares, WMS para mapas o XML para datos estructurados.

En cualquier caso, y recordando que los esfuerzos a corto plazo deberían centrarse en facilitar la *reutilización* directa e inmediata del mayor número posible de formatos y medios actuales, a la hora de decidir los formatos finales deberá buscarse siempre un equilibrio entre el potencial del formato actual de cara a la

[3] Resolución de 3 de octubre de 2012, de la Secretaría de Estado de Administraciones Públicas, por la que se aprueba la Norma Técnica de Interoperabilidad de Catálogo de estándares: http://www.boe.es/boe/dias/2012/10/31/pdfs/BOE-A-2012-13501.pdf

[4] Será necesario consultar con la asistencia técnica del área u organismo en cuestión o, en su defecto, con la coordinación general del proyecto de datos abiertos.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

reutilización de datos y el esfuerzo que puede suponer para el área u organismo la transformación de los datos a otros formatos más apropiados.

Nivel de detalle de los datos

Los datos deben proporcionarse con **el mayor nivel de detalle posible**, de forma que cada agente reutilizador pueda ser capaz de realizar con posterioridad el tratamiento de la información que considere necesario para sus fines específicos. Es por tanto importante **respetar el formato y detalle original de los datos** tal cual se generan en su fuente en la medida de lo posible, evitando cualquier modificación y alteración previa a su publicación incluso aunque el objetivo fuese facilitar la legibilidad de la información. La única excepción es el filtrado que pueda ser necesario realizar por cuestiones de **privacidad**, **confidencialidad** o **seguridad** como se ha comentado anteriormente.

De forma **complementaria** se podrían ofrecer también *agregaciones* de datos con el objetivo de facilitar la interpretación de la misma, pero nunca como sustituto de los datos *atómicos*, ya que de lo contrario podríamos estar impidiendo el acceso a ciertos datos que podrían ser relevantes para las necesidades de los *infomediarios*.

Calidad y fiabilidad

Los conjuntos de datos publicados deberán tener también un contenido razonablemente confiable y adecuado para ser directamente explotado por cualquier agente *reutilizador*, por lo que no deberían presentar problemas de **calidad** o **fiabilidad** más allá de lo razonable en cualquier caso. Si no es así, será necesario depurar previamente los datos antes de su publicación, determinando las medidas a tomar para llevar a cabo las reparaciones oportunas, con el objetivo de paliar los problemas que hayan sido previamente detectados.

Licencias de uso de los datos

Es recomendable que por defecto la publicación de cualquier conjunto de datos en el catálogo de datos, independientemente del área u organismo del que procedan, se realice según la modalidad general básica de puesta a disposición de los recursos de información reutilizable sin sujeción a condiciones específicas de acuerdo a lo establecido en el R.D. 1495/2011 y que incluye:

- La **condiciones generales** para la reutilización de los datos, como por ejemplo la prohibición de desnaturalizar el sentido de la información o la obligatoriedad de citar la fuente de los datos, entre otras.
- La **exclusión de responsabilidad** del organismo publicador en cualquier uso posterior que se vaya a realizar con los datos.
- La **responsabilidad del agente *reutilizador*** en el uso de los datos.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

- Información sobre cómo reflejar la **atribución** a la fuente original.

Si, excepcionalmente, un área u organismo optase por la aplicación de un régimen de licenciamiento alternativo para la puesta a disposición de los datos, es recomendable que lo haga exclusivamente a través de alguna de las licencias-tipo compatibles con los principios de los datos abiertos.

Con carácter general la licencia a aplicar para cualquier conjunto de datos dentro del proyecto Aragón Open Data será:

- Creative Commons-Reconocimiento 4.0[5] (CC-BY 4.0)

Con carácter excepcional[6] se podría aplicar alguna otra licencia compatible con los principios de los datos abiertos, como por ejemplo:

- Una licencia *Open Data Commons*[7]:
  - Public Domain Dedication and License (PDDL).
  - Attribution License (ODC-By).
  - Open Database License (ODC-ODbL).

4 – PUBLICACIÓN DE LOS DATOS

Finalmente el último paso consistirá en poner a disposición pública tanto los conjuntos de datos, como cualquier otra información adicional en forma de **metadatos** que pueda ser útil para su localización, clasificación y reutilización con el objetivo de que sean accesibles para el resto del mundo.

Los datos y metadatos serán recopilados por parte de los coordinadores designados en las distintas áreas y organismos, para posteriormente ser compartidos por algunos de los medios puestos a su disposición:

- Plantillas de metadatos.
- Tablas para la carga masiva de datos automatizada.
- Directamente a través de la aplicación del catálogo[8] accesible a través de http://opendata.aragon.es/ (próximamente).

[5] CC-BY 4.0: http://creativecommons.org/licenses/by/4.0/
[6] Ver caso de uso para los conjuntos de datos afectados por derechos de propiedad intelectual o industrial en el Anexo I.
[7] Open Data Commons. data/databases license: http://opendatacommons.org/licenses/
[8] El catálogo está basado en la plataforma Open Source CKAN http://ckan.org/, desarrollada por Open Knowledge, con modificaciones y adaptaciones realizadas por el Gobierno de Aragón y que están disponibles a través de: https://github.com/aragonopendata/Aragon-Open-data-Website.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

En caso de cualquier duda con respecto al procedimiento se recomienda contactar directamente con la coordinación general del proyecto a través del correo de soporte opendata@aragon.es.

De cara a su publicación los datos deberán estar disponibles o bien a través de un enlace directo a la fuente original de los datos o bien a través de algún tipo de **archivo** en cualquier formato que se actualizará de forma manual *subiendo* un nuevo archivo a la plataforma de forma periódica con la frecuencia que se haya establecido en cada caso (de esta forma también se mantendrá un archivo histórico de la evolución de los datos que puede resultar muy útil).

Metadatos

Los metadatos son **elementos fundamentales** a la hora de catalogar los conjuntos de datos, ya que nos permitirán clasificar y encontrar la información adecuadamente. Por tanto es muy importante no sólo proporcionar el mayor número de metadatos posible, sino hacerlo también de la forma adecuada siguiendo cuidadosamente las indicaciones que se dan en esta sección.

En caso de cualquier duda con respecto al significado o formato de alguno de los metadatos es conveniente consultar directamente con la coordinación general del proyecto [9] para evitar problemas que puedan afectar luego a la correcta reutilización de ese conjunto de datos.

Los metadatos se dividen en:

- **Obligatorios**: aquellos que, por su especial relevancia y también por cumplimiento con la legislación vigente, será obligado proporcionar.
- **Recomendados**: aquellos que, si bien no serían obligatorios para cumplir con la legislación vigente, igualmente es muy recomendable que se proporcionen por su relevancia a la hora de clasificar los datos adecuadamente.
- **Opcionales**: aquellos que, aún siendo igualmente recomendables, puede que no siempre estén disponibles.

Por otro lado, a su vez los distintos metadatos a proporcionar a veces corresponden a los **conjuntos de datos** y otras a las distintas **distribuciones**[10] que los componen.

TABLA RESUMEN METADATOS
CONJUNTOS DE DATOS

[9] Soporte disponible a través de opendata@aragon.es.
[10] Ver el apartado de definiciones al inicio del documento.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

OBLIGATORIOS	RECOMENDADOS	OPCIONALES
Título	Etiqueta(s)	Fecha actualización
Descripción	Licencia	Frecuencia actualización
Temática	Email del autor	Cobertura geográfica
Organismo		Ámbito temporal
		Validez
		Referencias adicionales
		Normativa
DISTRIBUCIONES		
OBLIGATORIOS	RECOMENDADOS	OPCIONALES
Distribución	Nombre	Información sobre formato
Formato	Descripción	

METADATOS OBLIGATORIOS

En este apartado se describen los metadatos mínimos que se deben proporcionar, tanto para cada uno de los conjuntos de datos que se vayan a publicar como para cada uno de los formatos de distribución de esos conjuntos.

CONJUNTOS DE DATOS

Título: Corresponderá al nombre representativo del conjunto de datos.

Se recomienda que sea un texto breve de no más de 10-12 palabras aproximadamente.

Descripción: Texto descriptivo del contenido del conjunto de datos y sus principales características.

Se recomienda incluir información sobre el tipo datos que contiene, su origen, posible utilidad, limitaciones, etc. aunque evitando replicar información que esté disponible a través de otros campos de metadatos.

La descripción debe ser lo más completa posible pero manteniendo una extensión razonable de no más de 3-4 párrafos. Cuando se trate de conjuntos de datos de cierta complejidad que necesiten una descripción más amplia lo recomendable es utilizar los metadatos de información adicional para añadir una referencia a algún tipo de guía o manual que pueda estar disponible.

Consultor: Carlos Iglesias – contact@carlosiglesias.es
Revisado por: Jose María Subero – jmsubero@aragon.es
David Portolés – informacion@idearium-consultores.com

Temática(s): Categoría o temática principal del conjunto de datos.

Se corresponderá con uno o varios valores (no se recomienda más de dos o tres a lo sumo) que deberemos seleccionar entre los disponibles en la clasificación oficial que se aplica en el catálogo de datos del Gobierno de Aragón:

Ciencia y tecnología	Energía	Seguridad
Comercio	Hacienda	Sociedad y bienestar
Cultura y ocio	Industria	Transporte
Demografía	Legislación y justicia	Turismo
Deporte	Medio ambiente	Urbanismo e infraestructuras
Economía	Medio rural y pesca	Vivienda
Educación	Salud	
Empleo	Sector público	

Es importante seleccionar las temáticas más adecuadas, que en principio no tendrían por qué estar directamente relacionadas con la estructura organizativa del Gobierno de Aragón ni de ninguna de sus áreas u organismos.

En el documento con la taxonomía del catálogo nacional[11] que se toma como referencia, se incluye también a modo de orientación una relación de los temas más comunes o posibles subcategorías que se corresponden con cada temática principal que puede ser de utilidad a la hora de seleccionar las temáticas.

Organismo: área u organismo responsable del conjunto de datos.

Se indicará el organismo responsable de los datos (ya sea un determinado Servicio, Unidad, Secretaría, Gabinete, Dirección, Subdirección, Departamento, Entidad…) dentro del organigrama oficial del Gobierno de Aragón[12].

Debe indicarse siempre el organismo con la mayor especificad posible (a nivel de servicio o similar). En caso de no estar claro, se indicaría como organismo responsable al propio Gobierno de Aragón.

DISTRIBUCIONES

Distribución: indicará la forma de localizar una distribución o recurso específico del conjunto de datos.

El fichero o recurso puede indicarse de una de las siguientes maneras:

[11] Taxonomía del Catálogo Nacional de Datos:
http://datos.gob.es/sites/default/files/files/12_tax_02.pdf
[12] Organigrama del Gobierno de Aragón:
http://servicios.aragon.es/organigrama_publico/PublicoServlet?accion=1

- *facilitando una dirección web (URL del tipo http://www.ejemplo.es) donde se puede acceder al conjunto de datos en cuestión (fichero, servicio, etc.)*
  ó
- *seleccionando un fichero desde nuestro propio equipo que contenga directamente los datos y se desee subir a la plataforma.*

Formato: indica el formato en que se encuentra representado el conjunto de datos.

El tipo de formato generalmente se identifica con un código 3 o 4 letras (CSV, XLS, HTML o JSON por ejemplo) y se suele corresponder con la extensión del archivo donde se guardan los datos. Algunos ejemplos de formatos reutilizables que se usan frecuentemente son:

***CSV**: para representar datos tabulares.*
***KML**: para representar información geográfica.*
***JSON**: para representar el intercambio de datos entre aplicaciones.*
***ODS**: para representar hojas de cálculo.*
***PX**: para representar datos estadísticos.*
***RDF**: para representar recursos semánticos.*
***RSS**: para representar la distribución de contenidos e información actualizada.*
***SHP**: para representar datos espaciales.*
***WMS**: para representar datos georeferenciados.*
***XLS**: para representar hojas de cálculo.*
***XML**: para representar vocabularios de datos personalizados.*

Es recomendable introducirlo siempre en mayúsculas para facilitar su legibilidad, y siempre se especificará el formato final de la distribución, nunca uno intermedio[13] (por ejemplo si se para acceder a una distribución de un conjunto de datos en formato CSV debemos hacerlo a través de una página HTML intermedia porque no tenemos un acceso directo el formato a indicar entonces será CSV, no HTML) .

Existe también un registro público completo que recoge todos los posibles tipos de formatos[14] existentes y que puede servir de ayuda, pero en caso de duda es recomendable ponerse en contacto con el coordinador general del proyecto Open Data[15].

METADATOS RECOMENDADOS

En este apartado se describen los metadatos adicionales recomendados que se deberían proporcionar siempre que estén disponibles, tanto para los conjuntos de datos como para cada uno de los formatos de distribución asociados.

CONJUNTOS DE DATOS

Etiqueta(s): consiste en una o varias etiquetas textuales específicas que servirán para clasificar el conjunto de datos.

---

[13] Ver caso de uso para los conjuntos de datos en formatos que requieran tratamiento previo en el Anexo I.
[14] IANA Media Types: http://www.iana.org/assignments/media-types/media-types.xhtml
[15] Soporte disponible a través del correo opendata@aragon.es

Es preferible que las etiquetas estén compuestas por una única palabra (o a lo sumo dos) y las distintas etiquetas se introducirán separadas entre sí por comas. Es recomendable limitarse a utilizar un número de etiquetas razonable para describir cada conjunto de datos, no más de 3 o 4 por ejemplo, priorizando aquellas que consideremos más relacionadas con el conjunto en cuestión y procurando también que se complementen con las temáticas principales que se han seleccionado previamente.

Licencia: nombre de las condiciones de uso o la licencia aplicable a los datos.

En general se utilizarán siempre los términos de uso para la "Reutilización de la información del sector público en el ámbito del sector público estatal" que ya han sido descritos en el apartado de licencias de uso de esta guía.

Con carácter general la licencia a aplicar para cualquier conjunto de datos dentro del proyecto Aragón Open Data será siempre:

- *Creative Commons-Reconocimiento 4.0[16] (CC-BY 4.0)*

Si en algún caso excepcional[17] fuese necesario aplicar otro tipo de licencia, debe consultarse primero con la coordinación general del proyecto a través del correo de consultas opendata@aragon.es.

Email del autor: información de contacto con el autor.

Se corresponderá con un correo electrónico general de contacto con el área u organismo que publica los datos, o el correo particular del responsable directo de los datos en cuestión.

DISTRIBUCIONES

Nombre: Corresponderá al título representativo de una distribución del conjunto de datos.

Se recomienda que sea un texto breve de no más de 10-12 palabras y que sirva para diferenciar la distribución en cuestión de cualquier otra, indicando por ejemplo las características o la forma de acceso que la diferencia del resto.

Descripción: Texto descriptivo del contenido de la distribución.

La descripción debe proporcionarse con una extensión razonable de no más de 2-3 párrafos. Se recomienda centrarse en cualquier información distintiva que sea particular de la distribución en cuestión.

[16] CC-BY 4.0: http://creativecommons.org/licenses/by/4.0/
[17] Ver por ejemplo el caso práctico de conjuntos de datos afectados por derechos de propiedad intelectual o industrial en el anexo de este mismo documento.

METADATOS OPCIONALES

En este apartado se describen otros metadatos adicionales que se pueden proporcionar para un conjunto de datos específico. Aunque no sea obligatorio cumplimentar los campos solicitados en este apartado, si su información es conocida o fácilmente averiguable, es muy recomendable facilitarla para mejorar la calidad y utilidad de los metadatos disponibles.

CONJUNTOS DE DATOS

Fecha de actualización: fecha de última actualización del conjunto de datos.

*Se indicará la fecha de última actualización realizada **al contenido del conjunto de datos** (no en los metadatos) siguiendo el formato AAAA-MM-DD, es decir, el año con un número de cuatro cifras, el mes con un número siempre de **dos cifras** (del 01 al 12) y el día con un número siempre de **dos cifras** (del 01 al 31).*

Por ejemplo: 2011-11-14 para el 14 de Noviembre de 2011.

Frecuencia de actualización: representa el periodo de tiempo estimado que transcurre entre cada actualización del conjunto de datos.

Se representará a través del valor más aproximado de entre las siguientes medidas (ordenadas de mayor a menor frecuencia):

Trienal; bienal; anual; semestral; cuatrimestral; trimestral; bimestral; mensual; bimensual; quincenal; trimensual; semanal; bisemanal; trisemanal, diaria, horaria e instantánea.

Por ejemplo:

- *Actualización todos los años: **anual***
- *Actualización cada 2 meses: **bimestral***
- *Actualización 2 veces al mes: **bimensual***
  *Actualización 2 veces a la semana: **bisemanal***

Ámbito geográfico: representa el espacio geográfico cubierto por los datos.

Si coincide con el ámbito general del Catálogo (Comunidad Autónoma de Aragón) no será necesario indicarlo, pero sí en el caso de que sea cualquier otro (provincia, municipio, comarca, etc.).

Debe indicarse siempre el ámbito geográfico con la mayor especificad aplicable[18] en cada caso (ya sea a nivel de provincia, comarca o municipio). En caso de no estar claro el ámbito de aplicación, se indicaría la Comunidad Autónoma de Aragón.

[18] El listado completo de comarcas y municipios puede consultarse en: http://idearagon.aragon.es/toponimia/comarca.htm

Ámbito temporal: identifica la fecha inicial y la fecha final del periodo de tiempo cubierto por el conjunto de datos.

Se trata de dos fechas correspondiendo al inicio y fin del periodo temporal cubierto por los datos en la forma AAAA-MM-DD, es decir, el año con un número de cuatro cifras, el mes con un número siempre de ***dos cifras*** *(del 01 al 12) y el día con un número siempre de* ***dos cifras*** *(del 01 al 31).*

Por ejemplo: inicio 2013-01-01 y fin 2013-12-31 sería el periodo desde el 1 de Enero de 2013 hasta el 31 de Diciembre de 2013.

Validez: identifica la fecha final de validez o *"caducidad"* de los datos a partir de la cual pueden perder su relevancia por cualquier motivo (modificación, actualización, etc.)

Se indicará la fecha de fin de validez de los datos siguiendo el formato AAAA-MM-DD, es decir, el año con un número de cuatro cifras, el mes con un número siempre de ***dos cifras*** *(del 01 al 12) y el día con un número siempre de* ***dos cifras*** *(del 01 al 31).*

Por ejemplo: 2014-01-31 indicaría que los datos serían válidos hasta 31 de Diciembre de 2014.

Referencias adicionales: direcciones web (URLs del tipo http://www.ejemplo.es) de páginas, documentos u otros recursos que contengan información ampliada sobre el conjunto de dato.

Por ejemplo, podría tratarse de un enlace a un documento en formato PDF dónde se explica la metodología para la recogida de esos datos o un enlace a una aplicación interactiva a través de la cual se explican los datos de una forma visual.

Normativa: direcciones web (URLs del tipo http://www.ejemplo.es) de páginas, documentos u otros recursos que contengan normativa relativa al conjunto de datos.

Por ejemplo podría enlazarse a una página web que muestre una ley u ordenanza que regule los datos contenidos en el conjunto.

DISTRIBUCIONES

Información sobre el formato: direcciones web (URLs del tipo http://www.ejemplo.es) de páginas, documentos u otros recursos que contengan información adicional, generalmente de carácter técnico, sobre los formatos utilizados por las distribuciones.

Por ejemplo, podría tratarse de un enlace a una página web donde puede encontrarse más información sobre el formato empleado, o un enlace a un esquema XSD asociado a un conjunto de datos en formato XML.

ANEXO I – EJEMPLO DE METADATOS

A continuación se muestra un ejemplo completo de cuáles serían los metadatos correspondientes a un hipotético conjunto de datos sobre convocatorias de empleo público que contase con dos distribuciones en distintos formatos.

Conjunto de Datos

Título: Convocatoria de empleo público del Servicio de Salud.

Descripción: Convocatorias de oposiciones y empleo público del servicio Aragonés de Salud publicados en el BOA a partir del 01/01/2014, incluyendo título de la convocatoria, número de plazas, temarios oficiales y plazos de los procedimientos.

Temática: Empleo; Sector Público

Etiqueta(s): oposiciones, convocatoria

Organismo: Servicio Aragonés de Salud

Email del autor: salud@aragon.es

Licencia: Creative Commons – By 4.0

Ámbito temporal: del 2014-01-01 al 2014-12-31

Frecuencia actualización: Semanal

Fecha actualización: 2014-06-28

Validez: 2014-12-31

Cobertura geográfica: Comunidad Autónoma de Aragón

Referencias adicionales:
http://aragon.es/Salud/Oposiciones/ListaAprobadosEnfermeria.pdf

Normativa: http://boa.aragon.es/oposiciones/bases-convocatoria=271213

Distribución 1

Nombre: Últimas convocatorias en JSON.

Descripción: Acceso a los principales datos de las tres últimas convocatorias durante el año en tiempo real a través del formato JSON.

Distribución:
http://boa.aragon.es/oposiciones/convocatoria=271213&format=JSON

Formato: JSON

Información sobre formato: http://json.org/

Distribución 2

Nombre: Histórico de convocatorias en XML.

Descripción: Acceso al histórico de todas las características de todas las convocatorias del año a través del formato XML.

Distribución:
http://boa.aragon.es/oposiciones/convocatoria=271213&format=XML

Formato: XML

Información sobre formato: http://boa.aragon.es/oposiciones/esquema.xml

ANEXO II – CASOS PRÁCTICOS

A continuación se dan algunas indicaciones a seguir en ciertos casos prácticos de publicación de conjuntos de datos que se dan con cierta frecuencia.

Conjuntos de datos afectados por derechos de propiedad intelectual o industrial

Cuando en la generación de datos y elaboración de documentos públicos intervienen otras entidades o empresas externas a las áreas y organismos del Gobierno de Aragón, en virtud de contratos o acuerdos de colaboración, deberá **comprobarse si cuentan con los derechos de propiedad intelectual o industrial suficientes** para permitir la reutilización efectiva por terceros de los correspondientes conjuntos de datos. En el caso de que no sea así, se puede optar por:

- Obtener del legítimo propietario la cesión de los derechos necesarios.

  ó

- No autorizar la reutilización de los conjuntos de datos afectados.

Si se opta por solicitar la cesión de derechos, y estos no fuesen conformes a los términos de uso generales previamente establecidos por el Gobierno de Aragón, se podría optar por establecer unas condiciones específicas de reutilización para ese conjunto de datos en concreto ajustadas a los términos de cesión de derechos finalmente obtenidos.

Por ejemplo, si cualquier área u organismo realizase una exposición con obras artísticas de distintos autores y quisiera publicar un conjunto de datos con fotografías y otros datos de todas esas obras, para poder hacerlo debería primero obtener el permiso de sus respectivos propietarios.

Conjuntos de datos con datos de carácter personal

La Ley 11/2007, de 22 de junio, de acceso electrónico de los ciudadanos a los servicios públicos, establece en su artículo 4.a) que el uso de tecnologías de la información debe contemplar el **respeto a la normativa sobre protección de datos de carácter personal y con los derechos al honor y a la intimidad**.

Por ello, para hacer compatible la publicación de conjuntos de datos con el respeto a la normativa establecida de forma previa a su publicación, se deberá **aplicar al**

conjunto de datos afectado un proceso, conocido como *disociación no reversible*, **que evite una posible identificación de personas a posteriori**. Por tanto, se recomienda con cada conjunto de datos afectado proceder de la siguiente manera:

1. Determinar cuáles son los elementos del conjunto de datos que podrían hacer posible la identificación de una persona: por ejemplo *nombre y apellidos*, *dirección postal* o *electrónica*, *teléfono*, *fax*, *DNI*, *Nº Seguridad Social*, *fotografía personal*, *firma*, *dirección IP*, etc.
2. Eliminar cualquier dato identificativo antes de la publicación del conjunto de datos.

En el caso de que la **disociación no fuese posible**, ya sea por dificultades técnicas, de coste u otras, **no se podrá publicar dicho conjunto**. Es importante también recordar que tanto las personas jurídicas como los empresarios individuales y los ficheros que se limiten a incorporar los datos profesionales de las personas físicas quedan excluidos del régimen de protección de datos de carácter personal[19].

Por ejemplo el padrón de habitantes debe contener obligatoriamente al menos los siguientes datos:

- *nombre y apellidos;*
- *sexo;*
- *domicilio habitual;*
- *nacionalidad;*
- *lugar y fecha de nacimiento;*
- *número de DNI o equivalente;*
- *nivel de estudios;*

Entre estos datos tanto el conjunto de nombre y apellidos como el DNI y el domicilio completo podrían hacer posible la identificación de una persona, y por lo tanto son datos que deberían filtrarse antes de la publicación.

Conjuntos de datos distribuidos en formatos propietarios

Aunque siempre se debe optar como primera opción por los **formatos abiertos** a la hora de publicar conjuntos de datos, también se podrán catalogar otras distribuciones en formatos propietarios, **siempre y cuando se trate al menos de un formato estructurado**.

Ejemplo de formatos propietarios estructurados bastante populares son XLS – formato de hoja de cálculo desarrollado por Microsoft – o SHP – formato de datos espaciales desarrollado por ESRI.

[19] Según Real Decreto 1720/2007 por el que se aprueba el Reglamento de desarrollo de la Ley Orgánica de protección de datos de carácter personal.

Por ejemplo, si tenemos un conjunto de datos disponible en formato XLS, podríamos publicarlo además fácilmente en formato CSV para que estuviese también disponible en un formato abierto.

Conjuntos de datos distribuidos en formatos no estructurados

En estos casos, como ocurre por ejemplo con muchos documentos de texto tradicionales (TXT, PDF, DOC, ODT), la única posibilidad de reutilización automatizada suele ser a través de procesamiento del lenguaje natural, pudiendo dar lugar en la mayoría de ocasiones a que finalmente los documentos no contengan información estructurada de utilidad, haciendo que los amplios recursos invertidos para el procesamiento automatizado sea inútiles.

Por tanto, este tipo de documentos **no deben considerarse prioritarios en el proceso de selección para la publicación**. Si los documentos estuviesen basados en datos que hayan sido recolectados para su elaboración, como ocurre frecuentemente con los informes de todo tipo, es recomendable publicar dichos datos de forma separada en el formato estructurado y reutilizable que se considere más apropiado en cada caso, acompañados en este caso del informe como documentación adicional.

Por ejemplo, si tenemos un informe en formato PDF sobre un estudio de movilidad en la región en el que se comentan numerosas estadísticas sobre tráfico, transporte, movilidad inter-urbana, etc. y queremos publicarlas para su reutilización, en lugar de simplemente publicar el informe en su formato actual deberíamos publicar esas estadísticas en un formato estructurado que fuese apto para su reutilización, como por ejemplo CSV, y luego podríamos incluir también el informe como documentación adicional.

Conjuntos de datos distribuidos en formatos de imagen, audio o vídeo

Esto es lo que sucede por ejemplo cuando los datos que se quieren compartir consisten en fotografías, videos o sonidos, como por ejemplo un mapa de bits con información cartográfica o un conjunto de fotos de zonas turísticas de la ciudad, cuyo formato natural suele ser de tipo imagen (como JPG, GIF, TIFF, etc.) o en forma de audio y/o video (MP3, WMV, MP4, etc.). En estos casos se procederá igual que con cualquier otro conjunto de datos, **proporcionando los metadatos necesarios a la hora de catalogarlos**.

No se debe confundir con otros casos en los que se utilicen las imágenes por ejemplo para capturas de pantallas de datos, exportaciones de datos en formato imagen o simplemente con documentos impresos que han sido escaneados. En estos casos, esos datos **no son directamente catalogables** ya que sería imposible reutilizarlos. Para su catalogación sería necesario que el área u organismo responsable **publicase dichos conjuntos de datos directamente en su formato estructurado original**, ya sea éste propietario o no.

Por ejemplo, podemos querer publicar el archivo fotográfico de la región, y en ese caso los datos se publicarán directamente en algún tipo de formato gráfico como JPG o similar, ya que ese es su formato "natural".

Sin embargo, si queremos publicar los presupuestos del Gobierno de Aragón por ejemplo, en este caso no sería adecuado escanear el documento correspondiente y publicarlo también como un JPG u otro formato gráfico, sino que habría que acceder a la fuente original de los datos para poder publicarlos en algún formato estructurado adecuado para ese tipo de información, como un CSV, una hoja de cálculo o similar.

Conjuntos de datos distribuidos en formatos que requieran tratamiento previo

En principio debe evitarse distribuir conjuntos de datos en formatos que **requieran de un tratamiento previo para poder ser utilizados**, como puede ser la necesidad de hacer un proceso de ***descompresión*** o facilitar una **contraseña** para el acceso a su contenido, ya que de este modo se estaría limitando su reutilización directa.

Entre los ejemplos del tipo de formatos que comúnmente pueden estar afectados por este tipo de limitación destacan ZIP, RAR y otros formatos de *compresión* – que no son realmente formatos finales de representación y además pueden fácilmente protegerse mediante contraseñas.

Debemos evitar siempre que sea posible el uso de este tipo de formatos comprimidos siguiendo algunas sencillas reglas:

- Si sólo vamos a publicar un único archivo de distribución es mejor no utilizar ningún formato de compresión, sino directamente el formato final.
- Si tenemos los datos distribuidos en varios archivos, sería conveniente agruparlo en un único archivo antes de su publicación (por ejemplo utilizando varías páginas en una misma hoja de cálculo en lugar de distintas hojas individuales).
- Si vamos a publicar el mismo conjunto en varios formatos distintos (por ejemplo XLS y JSON) cada uno de ellos se correspondería con una distribución independiente y por tanto deberían publicarse por separado y no en un mismo archivo comprimido.

Aún así, puede haber algunos casos en los que seguirá siendo necesario publicar varios archivos simultáneamente que se corresponden con una única distribución y que no es posible agrupar en un único archivo. En estos casos, se podría utilizar un formato de compresión para ello, pero en los metadatos de la distribución deberemos indicar como formato el que corresponde a la representación final, y no el que se ha utilizado únicamente en la encapsulación para su distribución comprimida.

Por ejemplo, podemos publicar datos sobre transporte público en formato GTFS que, según su propia especificación, se componen de una serie de archivos independientes, todos ellos en formato CSV, que contienen varios datos relacionados con aspectos como paradas, rutas, horarios, etc.

Según la especificación, estos archivos deberán estar recopilados en un archivo ZIP único, pero a la hora de proporcionar el metadato correspondiente al formato de estos archivos deberemos indicar que es GTFS (el formato de representación final) y no ZIP (el formato que se ha utilizado para encapsularlos).

Conjuntos de datos de acceso indirecto vía directorios, buscadores y otros

En ocasiones puede suceder que no haya posibilidad de acceder de forma directa a los conjuntos de datos, sino que previamente sea necesario **interactuar con una herramienta** de búsqueda y/o selección de datos (generalmente a través de una aplicación web) que devuelve los datos deseados como resultado (generalmente en formato HTML).

Si la herramienta ofrecida permite también la exportación de los datos en algún otro formato estructurado (por ejemplo CSV, XLS, etc.), debe priorizarse la catalogación en cualquiera de los **formatos estructurados de exportación ofrecidos**, siendo recomendable incluir también una referencia a la herramienta de origen como información adicional.

Por ejemplo, si tenemos una aplicación online del Gobierno de Aragón que nos permite acceder al directorio de centros de salud (por localización, provincia, municipio, localidad, etc.), en lugar de catalogar directamente esa página de la aplicación lo que deberíamos hacer es proporcionar esa misma información a través de formato estructurado, como por ejemplo un servicio web JSON o un fichero CSV, y catalogar ese conjunto de datos en su lugar, pudiendo incluir además una referencia a la aplicación como documentación o recurso adicional.

Conjuntos de datos no digitalizados

Cuando nos encontremos con conjuntos datos que han sido publicados por un área u organismo únicamente por **medios no electrónicos**, generalmente en forma de publicaciones impresas, será preciso localizar la fuente de datos original en las bases de datos y sistemas internos, a fin de poder extraer el conjunto de datos equivalente de cara a su publicación en formato electrónico.

Si dicha fuente no existiese o no fuera localizable, debe valorarse la posibilidad a medio plazo de **digitalizar** el contenido de dichas publicaciones en un **formato electrónico apropiado para facilitar su reutilización**.

Por ejemplo, si contamos con un buzón en las distintas dependencias de los órganos asociados al Gobierno de Aragón a través de los cuáles los ciudadanos pueden introducir sus sugerencias escritas en papel, además de mantener un archivador con esas sugerencias deberíamos almacenarlas también en algún tipo de base de datos u hoja de cálculo para que luego se pudiesen tratar automáticamente y de forma centralizada.